DOCUMENT RESUME

ED 283 837 TM 870 341

AUTHOR Mitchell, Karen J.; Anderson, Judith A.
TITLE Estimation of Interrater and Parallel Forms Reliability for the MCAT Essay.
PUB DATE Apr 87
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Essay Tests; Estimation (Mathematics); Generalizability Theory; Higher Education; Holistic Evaluation; *Interrater Reliability; Medical Education; *Test Length; Test Reliability; Timed Tests; *Writing Evaluation
IDENTIFIERS Medical College Admission Test; *Parallel Forms Reliability; Parallel Test Forms

ABSTRACT

The Association of American Medical Colleges is conducting research to develop, implement, and evaluate a Medical College Admission Test (MCAT) essay testing program. Essay administration in the spring and fall of 1985 and 1986 suggested that additional research was needed on the development of topics which elicit similar skills and meet standard criteria for overall difficulty. Testing of essay prompts differing in focus and length of response was undertaken with the Fall 1986 MCAT administration. Some examinees responded to a 45-minute and a 15-minute essay, while others wrote two 30-minute essays. The 45-minute and 30-minute topics led to papers of notably different quality levels. The 45-minute response period allowed for: (1) fuller treatment of issues and more complete argument development; (2) better demonstration of skills involved in organizing and presenting ideas and sustaining a line of reasoning; (3) testing of higher level writing and analysis skills, and (4) addressing the objectives of the MCAT writing assessment. The 30-minute prompts produced papers with simplistic argument statements and an increased number of surface errors. The Advisory Committee to the MCAT Essay Pilot Project judged concerns about the 30-minute essays to be too marked to promote the use of 60-minute forms and felt the 45-minute response period gave the examinees an important opportunity to plan, organize, and execute a thoughtful essay. (BAE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED283837

ESTIMATION OF INTERRATER AND PARALLEL FORMS RELIABILITY FOR THE MCAT ESSAY

Karen J. Mitchell, Ph.D.
Judith A. Anderson

Section for Student and Educational Programs
Association of American Medical Colleges, Suite 200
One Dupont Circle, N.W. Washington, D.C. 20036

Paper presented at Annual Meeting of the
American Educational Research Association,
Washington, D.C., 24 April 1987

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. Mitchell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 341

Estimation of Interrater and Parallel Forms Reliability for the MCAT Essay

The Association of American Medical Colleges is conducting research on the incremental validity associated with adding an essay to the Medical College Admission Test (MCAT). The objectives of the research are to develop, implement, and evaluate an MCAT essay testing program. Evaluation activities for the essay pilot project are described in "MCAT Essay Pilot Project Evaluation Plan" (AAMC, 4 March 1985). The analysis plan for initial stages of the pilot project are discussed in "Analysis Plan for the MCAT Essay Pilot Project" (AAMC, 17 April 1985).

The essay was administered on a pilot basis with the Spring and Fall 1985 and Spring and Fall 1986 MCAT administrations. Pilot administration of the essay will continue through 1987. Topics for the exam are designed to provide examinees with the opportunity to demonstrate skill in: 1) developing a central idea, 2) synthesizing concepts and ideas, 3) separating relevant from irrelevant information, 4) developing alternative hypotheses, 5) presenting ideas cohesively and logically, and 6) writing clearly, observing the accepted practices of grammar, syntax, punctuation, and spelling consistent with timed, first draft composition.

After each MCAT administration, experienced scorers meet to rate a sample of papers systematically selected to represent the demographic and academic characteristics of the examinee population. The rater training and scoring sessions are structured using White's (1985) suggestions for organizing and managing holistic essay readings. Essays are rated on a 6-point holistic scale by two readers. The two assigned ratings are summed to yield a final score ranging from 2 to 12. A small number of papers are rated by four readers to provide interrater reliability estimates.

Interrater reliability estimates for data from the first three administrations were calculated using procedures from generalizability theory. Coefficients ranged from .70 to .73. Score variation was partitioned into that due to subjects and that attributable to factors such as: 1) the particular readers who score the papers, 2) the tables at which the readers are grouped, 3) the batches to which the papers are assigned, 4) the days on which the papers are read, and/or 5) unexamined factors. The rank ordering of these variables in terms of their impact on score variation for past readings is batch, table, day by reader, day by batch, and reader.

Essay data from these administrations suggested that additional research was needed on the development of topics which elicit similar skills and meet standard criteria for overall difficulty. Mean differences between similar examinee groups on the first three testings called into question the equivalence of essay topics across administrations. Test results for small groups of MCAT re-examinees also suggested noncomparable difficulty levels or skill requirements for essay prompts. Because it is planned that unique essay questions will be presented at each MCAT administration, it is necessary to establish the equivalence of essay topics across different testings. Score equivalence is necessary both to provide admissions committees with a frame of reference for the data and to facilitate comparison of scores in applicant pools with test data from different MCAT administrations. More importantly, score equivalence is essential to a guarantee of fairness to applicants.

Testing of essay prompts differing in focus and length of response was undertaken with the Fall 1986 MCAT administration. Some examinees responded to a 45-minute essay and a 15-minute essay, while others wrote two 30-minute essays. The forms and data were structured to answer two questions:

- Could 45-minute questions be developed which were similar in difficulty levels and skill requirements across administrations; i.e., could topics be drafted to elicit the same mean scores and rank orderings of examinees across testings?

- Would the administration of a 60-minute form with two 30-minute topics lead to sufficient comparability and reliability of form data?

The rationale for development of the 60-minute forms was that administration of two topics theoretically leads to better representation of the skill domain tapped by a form and, hence, results in increased comparability and reliability of form data. MCAT essay data and the literature on writing assessment suggest that the particular content of an essay prompt heavily influences the quality of an examinee's response. The presentation of two different prompts in one form gives the examinee multiple opportunities to draft a response that best represents his or her ability.

Fall 1986 data revealed that each 45- and 30-minute topic elicited a wide range of response. There were no average score differences between examinees grouped by gender, rural/urban status, age, and number of years of postsecondary education. Very small differences in mean scores were noted for examinees grouped by racial/ethnic status; these differences were smaller, however, than those observed on the other MCAT tests. Correlations between the essay and other MCAT tests indicated that the essay was assessing a skill or skills unexamined by the other tests. The findings were similar to those of the first three administrations.

Mean score differences and test-retest reliabilities for the 45-minute topics, however, indicated that additional work is needed on the development of parallel 45-minute prompts. Mean scores for the different

45-minute topics ranged from 7.1 to 7.7. Test-retest reliabilities for Spring 1986/Fall 1986 repeaters ranged from .38-.58. These reliabilities are corrected for restriction in range.

Mean scores for the 60-minute forms ranged from 7.1 to 7.4. Reliability estimates for the 60-minute forms were based on generalizability coefficients. Test-retest reliability estimates were not available for the 60-minute forms because this test format was not administered prior to Fall 1986. The design for the generalizability study is depicted below.

FIGURE 1

ANOVA Design Used for Fall 1986

	D_1		D_2		D_3	
	F_1	F_2	F_3	F_4	F_5	F_6
	$B_1 \ldots B_3$	$B_4 \ldots B_6$	$B_7 \ldots B_9$	$B_{10} \ldots B_{12}$	$B_{13} \ldots B_{15}$	$B_{16} \ldots B_{18}$
$S_1 \ldots S_{20}$						$S_{341} \ldots S_{360}$
R_1						
R_2						
R_3						
R_4						

D = Duplication
F = Form
B = Batch
S = Subject
R = Reading (not Reader)

F_1 = Topics EG&GE
F_2 = Topics FH&HF
F_3 = Topics EF&FE
F_4 = Topics GH&HG
F_5 = Topics EH&HE
F_6 = Topics FG&GF

Four 30-minute topics were used. Each was paired with all others to produce the six forms. Order of presentation was counterbalanced. All topics were included in each duplication. Sixteen readers were randomly included in the study. Four readers were randomly assigned to each topic, i.e., Readers 1-4 read Topic E, Readers 5-8 read Topic F, etc. This was necessary for efficiency so that all readers did not need to be trained to score all topics. A duplication thus contained the input from all 16 readers.

Essays were randomly assigned to batches, and each batch contained twenty papers. Batches were designed so that the same people were grouped together for both topics within a form. That is, Batch 1 for Topic E contained papers from the same people as Batch 1 for Topic G. The dependent variable was the sum of two scores, one from the first topic and one from the second within a form. There were four scores of this type for each subject.

Two coefficients of differentiation were computed. These were the coefficients for subjects and forms. The coefficients for subjects included subject variance as true variation and the subject by reading interraction as error. The coefficient for forms included form variance as true variance and batch, subject, form by reading, batch by reading, and subject by reading as error. The data appear in Table 1. Test-retest reliability coefficients for the 60-minute forms can be expected to fall below the .81 coefficient for subjects but exceed those of the 45-minute forms.

TABLE 1

Generalizability Data for 60-Minute Forms

Sources	SS	DF	MS	F
Duplication	.878	2	.439	.06
Form (Dupl.)	22.119	3	7.373	1.37
Batch (Form, Dupl.)	64.356	12	5.363	.82
Subjects (Form, Dupl., Batch)	2089.962	342	6.111	9.52**
Reading	.348	3	.116	.04
Reading X Form (Dupl.)	12.789	9	1.421	1.36
Reading X Batch (Form, Dupl.)	37.656	36	1.046	1.63**
Dupl. X Reading	16.35	6	2.725	1.92
Reading X Subjects (Form, Dupl., Batch)	658.692	1026	.642	

Generalizability Coefficients:

Coefficient of Differentiation for Subjects = .81
Coefficient of Differentiation for Forms = .44

** $p<.05$

The Advisory Committee to the essay project and essay scorers observed that the Fall 1986 45- and Fall 1986 30-minute topics led to papers of notably different quality levels. It was the opinion of the Committee and the scorers that the provision of a 45-minute response period allowed for fuller treatment of issues and more complete argument development. They felt that the skills involved in organizing and presenting ideas and sustaining a line of reasoning were more appropriately demonstrated in a 45-minute than in a 30-minute time period. They judged that higher level writing and analysis skills were tested with the 45-minute paper, and these papers better addresed the objectives for MCAT writing assessment. The 30-minute prompts largely produced papers with

more simplistic argument statements and increased numbers of surface errors.

The highly skilled readers who score MCAT essays are trained to overlook the grammar, syntax and spelling errors that appear in first-draft writing. They are also trained to regard the depth of thought reflected in an essay in the context of the prompt itself and the time period allotted for response. The Committee was wary, however, of the admissions committee member's ability to factor these limiting conditions into his or her reading of an applicant's essay. The evaluation of a paper with incompletely or imperfectly developed logic or with a great number of surface errors may be colored by the layperson's regard for the importance of such characteristics. This problem may be exacerbated in contexts where assessors read one rather than both of the applicant's essays. The Essay Committee expressed concern that the provision of 30-minute papers would unjustifiably degrade the applications of many examinees. For many users surface problems with 30-minute essays would also call into question the grading scale and scoring scheme for the standardized essay.

In summary, the Advisory Committee to the MCAT Essay Pilot Project judged these concerns about the 30-minute essays to be too marked to promote the use of 60-minute forms. They felt that the 45-minute response period gave the examinees an important opportunity to plan, organize and execute a thoughtful essay. They judged this outcome to be a useful addition to the application process and they will continue to investigate procedures for enhancing the reliability of single-topic test.

References

Analysis Plan for the MCAT Essay Pilot Project (1985, April). Available from Section for Student and Educational Programs, Association of American Medical Colleges, Washington, D.C.

MCAT Essay Pilot Project Evaluation Plan (1985, March). Available from the Section for Student and Educational Programs, Association of American Medical Colleges, Washington, D.C.

Mitchell, Karen J. and Anderson, Judith (1986). Reliability of Holistic Scoring for the MCAT Essay. Educational and Psychological Measurement, 46,3.

White, Edward M. (1985). Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating and Improving Student Performance. San Francisco, California: Jossey-Bass Publishers.